Version dated: October 31, 2018

Exploring and Visualising Spaces of Tree Reconciliations

Katharina T. Huber[1], Vincent Moulton[1], Marie-France Sagot[2], and Blerina Sinaimeri[2]

[1] *School of Computing Sciences, University of East Anglia, Norwich, United Kingdom*

[2] *Inria Grenoble - Rhône-Alpes; Inovallée 655, avenue de l'Europe, Montbonnot, 38334 Saint Ismier cedex, France, and Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558; 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France*

Corresponding author: Vincent Moulton, School of Computing Sciences, University of East Anglia, Norwich, United Kingdom; E-mail: v.moulton@uea.ac.uk.

Abstract.— Tree reconciliation is the mathematical tool that is used to investigate the coevolution of organisms, such as hosts and parasites. A common approach to tree reconciliation involves specifying a model that assigns costs to certain events, such as cospeciation, and then tries to find a mapping between two specified phylogenetic trees which minimises the total cost of the implied events. For such models, it has been shown that there may be a huge number of optimal solutions, or at least solutions that are close to optimal. It is therefore of interest to be able to systematically compare and visualise whole collections of reconciliations between a specified pair of trees. In this paper, we consider various metrics on the set of all possible reconciliations between a pair of trees, some that have been defined before but also new metrics that we shall propose. We show

that the diameter for the resulting spaces of reconciliations can in some cases be determined theoretically, information that we use to normalise and compare properties of the metrics. We also implement the metrics and compare their behaviour on several host parasite datasets, including the shapes of their distributions. In addition, we show that in combination with multidimensional scaling, the metrics can be useful for visualising large collections of reconciliations, much in the same way as phylogenetic tree metrics can be used to explore collections of phylogenetic trees.

Implementations of the metrics can be downloaded from:

`https://team.inria.fr/erable/en/team-members/blerina-sinaimeri/reconciliation-distances/`

(Keywords: reconciliation, reconciliation space, coevolution, phylogenetic tree)

Phylogenetic tree reconciliation is commonly applied to investigate the coevolution of various "entities" such as genes and species or hosts and parasites. It has been used to understand the relationship between gene and species evolution (Bansal *et al.* 2012; Doyon *et al.* 2011; Tofigh *et al.* 2011) the coevolution of parasites and their hosts (Charleston 1998, 2003), and the connection between species evolution and habitat change (Page 1994) The literature on phylogenetic tree reconciliation is extensive; see (Charleston 2003; Doyon *et al.* 2011; Eulenstein *et al.* 2010; Nakhleh 2013; Page and Charleston 1998; Ronquist 2002) for reviews and Bansal and Alm (2012); Charleston (1998); Conow *et al.* (2010); Merkle and Middendorf (2005); Page (1994); Ronquist (1995); Stolzer *et al.* (2012) for some key references in this area. In this paper, we focus on host-parasite coevolution. The approaches we propose could however be applied to any one of the other aforementioned scenarios.

Informally put, phylogenetic tree reconciliation is the problem of finding a mapping from one such tree to another. In our case, it is the parasite tree, that we denote by P, that is mapped to the other, the host tree, which we denote by H. Besides P and H, we are also given as input a map from the leaves of P to the leaves of H reflecting which parasites currently inhabit which hosts, and the aim is to find a map, or *reconciliation*, which represents the following events: cospeciation (when host and parasite speciate together), duplication (when the parasite speciates but not the host, both new parasite species remaining associated with the host), loss (when, for example, the host speciates but not the parasite, leading to the loss of the parasite in one of the two new host species or extinction or failure to sample the parasite lineage), and host switch (when the parasite speciates, one species remaining with its current host while the other switches to another). In the context of gene-species studies, this model is known as the DTL (for "Duplication, Transfer, and Loss") model for reconciliation (or DL model if host switches are not allowed). Reconciliation has been extensively studied within that model (see, for example,

Bansal *et al.* (2012); Doyon *et al.* (2011); Stolzer *et al.* (2012); Tofigh *et al.* (2011)), and also explored within the host-parasite model for example (Donati *et al.* 2015)).

Most approaches to tree reconciliation adopt a parsimony approach: given the assignment of a cost to each of the events, the solution sought is the one which minimises the total cost over all possible maps. When time consistency is ignored, that is, when we do not care whether the optimal mapping may involve one or more jumps back in time, the problem may be addressed using dynamic programming. Most often however, there is not a single such solution (Bansal *et al.* 2013; Chan *et al.* 2015; Donati *et al.* 2015; Doyon *et al.* 2009, 2012; Wu and Zhang 2012). Indeed, in many cases the number can be huge. Furthermore, depending on the costs assigned to each event, the optimal solutions may diverge greatly in terms of mapping, and importantly, also in terms of the number of each event in the solutions. This is observed even for the costs usually adopted in the literature, as was shown notably in (Donati *et al.* 2015) for the host-parasite model. In order to reach greater insight on the possible coevolution of the two sets of organisms, it is therefore important to consider all reconciliations, or collections thereof (Bansal *et al.* 2013; Chan *et al.* 2015; Donati *et al.* 2015; Doyon *et al.* 2009, 2012; Wu and Zhang 2012).

This situation is analogous to the one experienced when building phylogenetic trees where multiple solutions are generated and sometimes even suboptimal ones are considered (Hillis *et al.* 2005). What is different in relation to tree building, is that in the case of tree reconciliation, the existing methods avoid the NP-completeness of the problem by not checking for time consistency. In Donati *et al.* (2015) it was shown that this is often not a problem because, among all the optimal reconciliations, some may be time consistent, which further justifies the need to generate all of them. Notice that this may however not always resolve this issue, and in this case finding (all) suboptimal reconciliations may still provide a way of identifying time consistent optimal reconciliations. For the sake of simplicity, in this paper, we concentrate on considering all possible reconciliations, calling

however attention to the fact that the metrics considered could apply to both of the aforementioned contexts.

To better understand collections of reconciliations, it is natural to consider them within *spaces*. These are defined by taking the set of *all possible* reconciliations relative to a fixed model, and defining a metric on this set which is used to compare reconciliations. Such spaces allow for quantitative analyses to be performed, for example, to understand how suboptimal reconciliations behave (Doyon *et al.* 2009). Moreover, taking this holistic view point can be helpful to understanding how costs can effect the behaviour of reconciliations, as demonstrated in a recent work on "event-cost" spaces (Libeskind-Hadas *et al.* 2014). Considering spaces of reconciliations is analogous to the case for phylogenetic trees, where tree-spaces have also proven to be a useful tool for analysing the behaviour of likelihood tree optimisation models (Billera *et al.* 2001; Hillis *et al.* 2005; Jombart *et al.* 2017; Kendall and Colijn 2016).

When defining spaces of reconciliations, the choice of metric is crucial. Indeed, different metrics may pick out different features, or certain metrics may have systematic biases. Furthermore, from a practical point of view, some metrics are efficiently computable whereas others are not. The above is true in more general contexts, see for example the pros and cons of defining metrics on phylogenetic trees (Jombart *et al.* 2017; Steel and Penny 1993) and RNA structures (Moulton *et al.* 2000). For tree reconciliation spaces, edit distances have been the most commonly used metrics to date (see below). For example Doyon *et al.* (2009) used the edit distance to analyse suboptimal reconciliations under the DT model, and more recently edit distances were defined within a theoretical analysis for the DTL model (Chan *et al.* 2015). However, the edit distance is probably computationally hard to compute for the DTL model (see also Chan *et al.* (2015)). Moreover, for the reasons mentioned, it remains of importance to try other metrics.

In this paper, we define and study the relationship between different classes of

metrics defined on the set of all possible reconciliations relative to the DTL model presented in Tofigh *et al.* (2011) some of which have been already used in the literature. This is the case for the so-called *edit based* classes; the others are new ones that we introduce in this paper, which we call *function-based* and *tree-based* respectively. We have chosen a broad range of metrics, focusing on those that can be analysed from a theoretical perspective and can be efficiently computed. This makes them useful from a practical perspective, although our choice is not intended to be definitive in nature.

Besides showing that the three classes of metrics that we consider are indeed metrics and discussing the relationships between them, we theoretically determine the diameters (i.e., the maximum value that they can take over all possible pairs of reconciliations) for some of the metrics. This is important within practical applications where it is useful to normalise data for comparison purposes. Through an extensive set of computational experiments, we then study the distributions of the metrics that we have defined, and investigate in more depth how they are related to one another. Finally, we use the metrics together with multidimensional scaling to visualise collections of reconciliations, an approach that has proven successful for exploring phylogenetic tree-spaces (Hillis *et al.* 2005).

Preliminaries

Given a set X of taxa of size at least three, a *phylogenetic tree T (on X)* is a rooted tree which has a root vertex ρ_T with indegree zero (i.e., no edge coming into it) and outdegree at least two (i.e., at least two edges that are starting at it) and leaf set X. The tree T is *binary* if the root has two outgoing edges and, when directing all remaining edges away from it, every vertex of T that is not the root or a leaf has indegree one and outdegree two. We denote the vertex set of T by $V(T)$, the leaf set of T by $L(T)$ and we

let $V^o(T) = V(T) - L(T)$ denote the set of *interior* vertices of T. If $v \in V^o(T)$, we let $Ch(v)$ denote the set of *children* of v, and if $v \in V(T) - \{\rho_T\}$ we let $par(v)$ denote the *parent* of v in T.

We denote the partial order on $V(T)$ given by considering ancestors in T by $\succeq_T$ (in case the context is clear, we just use $\succeq$). For vertices $x, y \in V(T)$ with $(x \succ y)$ $x \succeq y$ we say that y is *(strictly) below* x (i.e., y is a (strict) descendant of x) and that x is *(strictly) above* y (i.e., x is an (strict) ancestor of y). Note that there exist vertices in T that are neither above nor below each other. If Y is a subset of $L(T)$ of size at least two, we let $lca_T(Y) = lca(Y)$ denote the *least common ancestor of the set* Y, that is, the lowest vertex in T which is above every element of Y.

Now, given phylogenetic trees P and H and a leaf map $\phi : L(P) \to L(H)$, a *reconciliation (with respect to* (P, H, ϕ)*)* is a map $\psi : V(P) \to V(H)$ which satisfies the following conditions:

(1) The map ψ restricted to the leaf set of P is equal to ϕ.

(2) If $v \in V^o(P)$, then

  (a) there exists no $v' \in Ch(v)$ such that $\psi(v')$ is a strict ancestor of $\psi(v)$.

  (b) there exists some $v' \in Ch(v)$ such that $\psi(v')$ is below $\psi(v)$.

This definition models reconciliations where cospeciations, duplications and host switches are all allowed (Tofigh *et al.* 2011). If we replace (2) by

(2') If $v \in V^o(P)$, then $\psi(v)$ is an ancestor $\psi(v')$ for all $v' \in Ch(v)$,

then we model reconciliations without host switches.

As an illustration of some of these concepts, consider the example presented in Fig. 1. In this example, P is a binary parasite tree whose leaves are $a, \ldots, e$ and whose interior vertices are α, β, γ and ρ_P. The children of γ are α and β and the parent of γ is

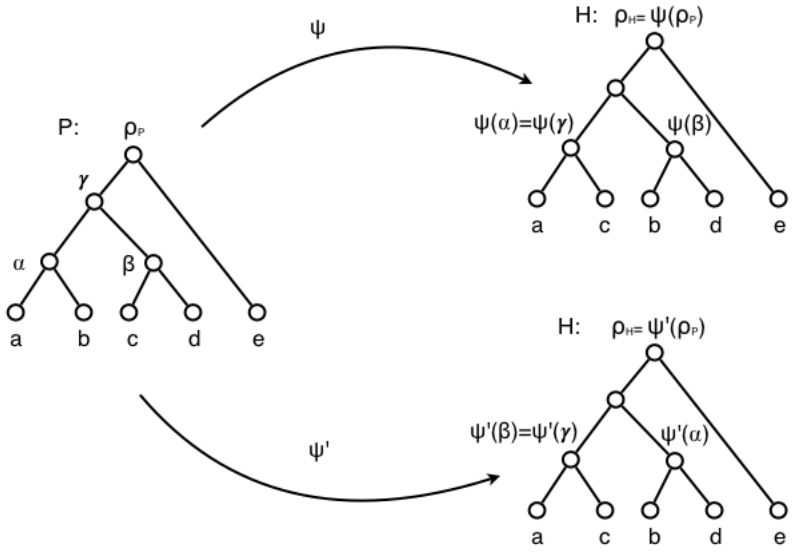


Figure 1: For the trees P and H on $X = \{a, b, c, d, e\}$, the two reconciliations denoted by ψ and ψ'. The map ϕ takes each leaf of P to the leaf of H with the same label.

ρ_P. Since γ is an ancestor of α we have $\alpha \preceq \gamma$. In fact, we have $\gamma \prec \alpha$ as α and γ are distinct vertices. Note that α is neither below nor above β. The least common ancestor of $a, \ldots, d$ is γ. Both maps ψ and ψ' are reconciliations of P with the host tree H where, for ease of readability, we have labelled the interior vertices of H with the images of the vertices of P under each reconciliation. So, for example, the image of β under ψ' is the least common ancestor of a and c in H. Note that both reconciliations satisfy Properties (2a) and (2b) but not Property (2'). We denote the set of reconciliations with respect to (P, H, ϕ) with host switches (no restriction) by $\mathcal{C}(P, H, \phi)$, and the set without host switches by $\mathcal{R}(P, H, \phi)$, so that in particular, $\mathcal{R}(P, H, \phi)$ is a subset of $\mathcal{C}(P, H, \phi)$.

We now state a basic result concerning reconciliations. For $v \in V(P)$, we let $m(v)$ denote the vertex in $V(H)$ given by

$$m(v) = lca_H(\{\phi(x) \,:\, x \in L(P) \text{ and } v \succeq_P x\}),$$

we let $A(v)$ be the subset of $V(H)$ given by

$$A(v) = \{u \in V(H) \,:\, \rho_H \succeq_H u \succeq_H m(v)\},$$

and we let $A^*(v)$ be the union of $A(v)$ and the set of vertices below $m(v)$ that are in the subtree of H spanned by the image of ϕ and $m(v)$. Informally speaking, $m(v)$ is the lowest vertex in H which v can be mapped to by a reconciliation and $A(v)$ comprises all vertices that are ancestors of $m(v)$. For the example of the host tree in Fig. 1, the vertex $m(\alpha)$ is the unique interior vertex in H that is not labelled and $A(\alpha)$ consists of that vertex and the root of H; $A^*(\alpha)$ consists of $A(\alpha)$ together with all vertices below $m(\alpha)$. The proof of the following lemma may be found in Supplementary Material I which is available in Dryad as https://doi.org/10.5061/dryad.g38g40b. Note that analogous observations have been made in, for example, (Doyon *et al.* 2009) and (Chan *et al.* 2015) but using different models for reconciliations.

Lemma 1. *If $\psi \in \mathcal{C}(P, H, \phi)$, then for all $v \in V(P)$ we have $\psi(v) \in A^*(v)$. Moreover, if $\psi \in \mathcal{R}(P, H, \phi)$, then for all $v \in V(P)$ we have $\psi(v) \in A(v)$.*

Metrics

In this section, we shall define various metrics on the set $\mathcal{C}(P, H, \phi)$ which can be restricted to also give metrics on $\mathcal{R}(P, H, \phi)$. These will be based on three different points of view: Considering reconciliations as functions (function based), representing reconciliations as trees (tree based), and defining operations for converting one reconciliation into another (edit based). The edit based approach has been considered in the literature before (see, for example, (Chan *et al.* 2015; Doyon *et al.* 2009)), and approaches that use events to measure the difference between reconciliations have also been

considered (Nguyen *et al.* 2013). These approaches use different models for reconciliations to the ones used in this paper.

Before proceeding, we recall that a *metric* on a set Y is a map D from $Y \times Y$ to the non-negative real-numbers such that

(i) for $y, y' \in Y$, $D(y, y') = 0$ if and only if $y = y'$;

(i) for $y, y' \in Y$, $D(y, y') = D(y', y)$ (symmetry);

(i) for $y, y', y'' \in Y$, $D(y, y'') \leq D(y, y') + D(y', y'')$ (the triangle inequality).

Informally put, a metric on a set Y allows one to assign a distance $D(y, y')$ between any pair of elements y and y' in Y which measures the dissimilarity of y and y'.

Function-based metrics

We begin by noting that if Y, Z are finite sets, $F(Y, Z)$ denotes the set of functions from Y to Z, and D is a metric on Z, then the distance measure d_D given by taking

$$d_D(f, f') = \sum_{y \in Y} D(f(y), f'(y))$$

for all $f, f' \in F(Y, Z)$ is a metric on the set $F(Y, Z)$. To illustrate this metric, assume that both Y and Z are the set $\{1, 2, 3\}$, that $f(1) = 2 = f'(3)$, $f(2) = 1 = f'(2)$, and $f(3) = 3 = f'(1)$. In addition, let D be the metric defined by putting $D(x, y) = |x - y|$, for all $x, y \in Z$. Then $D(f(1), f'(1)) = |f(1) - f'(1)| = |2 - 3| = 1$. Moreover $d_D(f, f') = \sum_{y \in \{1,2,3\}} D(f(y), f'(y)) = |f(1) - f'(1)| + |f(2) - f'(2)| + |f(3) - f'(3)| = 1 + 0 + 1 = 2$.

Using this observation it is now straightforward to define metrics on $\mathcal{C}(P, H, \phi)$ by considering different metrics D on the host tree. In particular, given a metric D on $V(H)$

we define the distance d_D between ψ, ψ' in $\mathcal{C}(P, H, \phi)$ by setting

$$d_D(\psi, \psi') = \sum_{v \in V(P)} D(\psi(v), \psi'(v)).$$

that is, $d_D(\psi, \psi')$ is the sum of the dissimilarity of $\psi(v)$ and $\psi'(v)$ under D, where the sum is taken over all vertices v of the parasite tree. Note that, by the above observation, d_D is a metric on $\mathcal{C}(P, H, \phi)$, i.e., the set of reconciliations where host switches are allowed.

In what follows, we shall consider two metrics D on H, namely (i) the discrete distance which measures the distance between any two vertices v and w in the host tree by taking that distance to be 0 if v and w are equal and 1 else and (ii) the path distance D_H which measures the distance between v and w by taking that distance to be the length of the (undirected) path in the host tree between v and w. For example, the discrete distance between the vertices $\psi(\alpha)$ and $\psi(\beta)$ in Fig. 1 is 1 whereas the path distance between them is 2. We denote the metric d_D that we obtain in these cases by d_{disc} (the discrete distance) and d_{path} (the path distance), respectively. Note that the path distance has been implicitly considered in Doyon *et al.* (2009) under the DL model.

Tree-based metrics

We now consider a class of metrics which is defined by representing reconciliations using phylogenetic trees. Given $\psi \in \mathcal{C}(P, H, \phi)$ we define a phylogenetic tree $H(\psi)$ with leaf set $L(H) \cup V(P)$ as follows. Starting with H, for each interior vertex w of H, we introduce a pendant edge at w with leaf labelled by v, for each vertex v of P with $\psi(v) = w$, and for each leaf w of H, we replace w with a new vertex that is adjacent to pendant edges labelled by w and by v, for each vertex v of P with $\psi(v) = w$. In Fig. 2, we illustrate this construction. Note that trees have been used to represent reconciliations in Scornavacca *et al.* (2013), but in a different way.

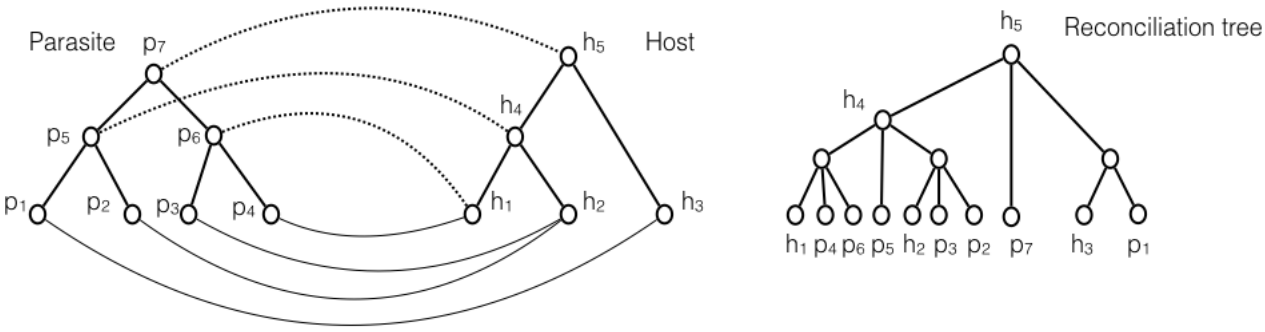


Figure 2: An example illustrating the construction of a phylogenetic tree from a reconciliation as described in the text.

The key observation used to obtain metrics on $\mathcal{C}(P, H, \phi)$ is that for $\psi, \psi' \in \mathcal{C}(P, H, \phi)$, the tree $H(\psi)$ is isomorphic as a phylogenetic tree to $H(\psi')$ if and only if $\psi = \psi'$. Hence, if D is a metric on the set of phylogenetic trees with leaf-set $L(H) \cup V(P)$, and for $\psi, \psi' \in \mathcal{C}(P, H, \phi)$, we let

$$d^D(\psi, \psi') = D(H(\psi), H(\psi')),$$

then it follows that d^D is a metric on $\mathcal{C}(P, H, \phi)$.

Various metrics have been defined for comparing phylogenetic trees (see for example Steel and Penny (1993)). Here we shall focus on two of these that are commonly used in the literature: The Robinson-Foulds distance and the triplet distance, which have the advantage of both being efficiently computable. By the above observation, by taking D to be either of these two metrics we obtain a metric on $\mathcal{C}(P, H, \phi)$ for which we shall use the same name. We denote them by d_{RF} and d_{tr}, respectively.

For completeness, we now briefly recall the definitions of the Robinson-Foulds distance and the triplet distance. If T is a phylogenetic tree, and v is a vertex of P, then we let T_v denote the subtree of T whose vertices are all below v. We let $L(v)$ denote the

leaf set of T_v, and we refer to $L(v)$ as the *cluster* in T associated to v (if v is the root or a leaf of T then we call $L(v)$ a *trivial* cluster). In addition, a binary phylogenetic tree on three leaves x, y, z is called a *triplet*; if for such a triplet the least common ancestor of x and y is strictly below its root, then the triplet is said to be in T if x, y, and z are leaves of T and $lca_T(x, y) \neq lca_T(x, z) = lca_T(y, z)$. The *Robinson-Foulds (respectively triplet) distance* between two phylogenetic trees is then defined by taking one-half of the number of clusters (respectively triplets) that are in one tree but not in the other.

Edit-based metrics

We now focus on a class of metrics that is defined in terms of converting one reconciliation into another using a minimal number of some specified operations. We do this using two maps which we now define. For a reconciliation ψ and a pre-given vertex v of P, the purpose of the first map, denoted by ψ_v^U, is to move $\psi(v)$ up by a vertex in H. In contrast, the purpose of the second map, denoted by $\psi_v^{Do,m}$, is to move $\psi(v)$ down to a pre-given child m of $\psi(v)$ in H. More formally, let $\psi \in \mathcal{C}(H, P, \phi)$. If v is an interior vertex of P and $\psi(v) \neq \rho_H$, then we define the map $\psi_v^U : V(P) \to V(H)$ by putting

$$\psi_v^U(w) = \begin{cases} \psi(w) & \text{if } w \neq v \\ par(\psi(v)) & \text{else.} \end{cases}$$

If v is an interior vertex of P that is not mapped to a leaf of H by ψ and $m \in Ch(\psi(v))$, then we define the map $\psi_v^{Do,m} : V(P) \to V(H)$ by putting

$$\psi_v^{Do,m}(w) = \begin{cases} \psi(w) & \text{if } w \neq v \\ m & \text{else.} \end{cases}$$

In other words, the maps ψ and ψ_v^U are the same except for the vertex v which is mapped

by ψ_v^U to the parent of the vertex $\psi(v)$ in the host tree. Similarly, the maps ψ and $\psi_v^{Do,m}$ only differ in v which is mapped by $\psi_v^{Do,m}$ to the specified child m of $\psi(v)$.

We say that the map ψ_v^U is obtained by applying an *up-operation* and $\psi_v^{Do,m}$ by applying a *down-operation* (cf. Chan *et al.* (2015); Doyon *et al.* (2009) for similar concepts). Note that the maps ψ_v^U and $\psi_v^{Do,m}$ are not necessarily contained in $\mathcal{C}(H, P, \phi)$ for any choice of v and m (see Supplementary Material I for necessary and sufficient conditions for this to be the case).

It is straightforward to see that the up/down-operations are mutual inverses. More specifically, if $\psi(v) \neq \rho_H$, then $(\psi_v^U)_v^{Do,\psi(v)} = \psi$, and if $\psi(v)$ is not a leaf of H and $m \in Ch(\psi(v))$ then $(\psi_v^{Do,m})_v^U = \psi$. Based on this observation, we define the *edit distance* $d_{ed}(\psi, \psi')$ between ψ and ψ' in $\mathcal{C}(P, H, \phi)$ to be the minimum number of up/down-operations that need to be applied, starting with ψ, to obtain ψ' (or vice-versa). In Supplementary Material I, we show that the edit distance is a metric on $\mathcal{C}(P, H, \phi)$; our proof works by viewing the edit distance as the metric on a certain graph that can be associated to $\mathcal{C}(P, H, \phi)$, and is similar to a comparable result that has been proven to hold for a different model of DTL reconciliation in Chan *et al.* (2015). In Supplementary Material I, we also show that when the up/down-operations can be applied to reconciliations in the set $\mathcal{R}(H, P, \phi)$ they give rise to new reconciliations in this set (i.e., they do not introduce host switches in such reconciliations), and that the resulting edit distance between any pair of reconciliations in $\mathcal{R}(H, P, \phi)$ is equal to their edit distance in $\mathcal{C}(P, H, \phi)$.

Dominance relations

Given a pair of metrics D, D' on the same set Y, we say that D *dominates* D' if for all $y, y' \in Y$ the distance between y and y' under D' is no larger then the distance between

them under D. Various domination relationships hold between the metrics that we have defined on $\mathcal{C}(P, H, \phi)$ which we now summarise.

First, note that when viewed as metrics on phylogenetic trees, the discrete distance is dominated by the path distance as the length of a path between any pair of distinct vertices in a tree is at least 1. Hence, the discrete distance d_{disc} is dominated by the path distance d_{path} on $\mathcal{C}(P, H, \phi)$. Also, note that the triplet distance dominates the Robinson-Foulds distance for phylogenetic trees (Huber *et al.* 2011, Lemma 1). Hence, on $\mathcal{C}(P, H, \phi)$, it follows that the Robinson-Foulds distance d_{RF} is dominated by the triplet distance d_{tr}.

Interestingly, the path distance is also dominated by the edit distance on $\mathcal{C}(P, H, \phi)$. Indeed, this follows since for any interior vertex v of P, we need to apply at least $D_H(\psi(v), \psi'(v))$ up/down-operations to ψ so that the map obtained by applying this sequence of operations assigns the vertex $\psi'(v)$ to v. However, in general the path and edit distances are different on $\mathcal{C}(P, H, \phi)$; see for example in Fig. 1 where the path distance between ψ and ψ' is 4 whereas the edit distance between ψ and ψ' is 6 (although the path and edit distances are in fact equal when restricted to $\mathcal{R}(H, P, \phi)$ – see for example Huber *et al.* (2018)). It would be interesting to know if there are any further dominance relationships between any of the discrete/path/edit distance and either the Robinson-Foulds distance or the triplet distance.

Diameters

In this section, we consider the problem of determining the diameter of the various metrics that we have defined on the sets $\mathcal{C}(P, H, \phi)$ and $\mathcal{R}(P, H, \phi)$. Knowing this quantity is useful for normalising metrics which, in turn, is useful for comparison purposes. The *diameter* of a metric D on a finite set Y is its maximum value taken over all pairs in Y; we

denote this quantity by $diam(Y, D)$. Note that diameters of reconciliation spaces under edit distances have been studied empirically under the DL model in Doyon *et al.* (2009), and more recently an algorithm was presented in Haack *et al.* (2018) for computing diameters of the event-based distance defined in Nguyen *et al.* (2013) under the DLT model.

We will first determine the diameter of the discrete and Robinson-Foulds distances on $\mathcal{C}(P, H, \phi)$. To do this we introduce some additional concepts. A *path system in* P is a set Λ of vertex-disjoint, directed paths in P which covers P (i.e., every vertex in P is contained in some path of Λ), and for which every path contains some leaf of P. It is straightforward to see that for any P there exists some path system, and that any path system of P has as many elements as $L(P)$. Moreover, given any path system Λ in P there always also exists a path system Λ^T in P which can be obtained by deleting all edges from the paths in Λ. Using again the example in Fig. 1, the single vertex paths a and c and the paths with vertices α b, vertices γ, β, d, and vertices ρ_P, e form a path system for the depicted parasite tree.

Now, given a path system Λ in P we define a map $\psi_\Lambda : V(P) \to V(H)$ by setting $\psi_\Lambda = \phi$ on the leaves of P and, for all interior vertices v of P, setting $\psi_\Lambda(v) = \phi(x)$, for x a leaf of P with v and x contained in the same path in Λ. It is straightforward to check that ψ_Λ is a reconciliation in $\mathcal{C}(P, H, \phi)$. We now give diameter bounds for the triplet distance and the Robinson-Foulds distance; for technical reasons we restrict ourselves to the case where ϕ is surjective (i.e., every host contains some parasite) as otherwise we could remove all hosts from H that do not contain a parasite (suppressing vertices with indegree and outdegree one, and identifying the root of the host tree with its unique child in case this has rendered it a vertex with outdegree one) without affecting the reconciliation under consideration. Note that in the following the map ψ_{root} is the reconciliation in $\mathcal{C}(P, H, \phi)$ defined by setting $\psi_{root}(v) = \phi(v)$ for all leaves of P and $\psi_{root}(v) = \rho_H$ for all interior vertices v of P.

Theorem 2. *Given (P, H, ϕ) with ϕ surjective, and any path system Λ in P, we have*

(i) $diam(\mathcal{C}(P, H, \phi), d_{disc}) = d_{disc}(\psi_{root}, \psi_\Lambda) = d_{disc}(\psi_{root}, \psi_{\Lambda^T}) = |V^o(P)|$.

(ii) $diam(\mathcal{C}(P, H, \phi), d_{RF}) = d_{RF}(\psi_\Lambda, \psi_{\Lambda^T}) = 2(|V(H)| - 1)$.

Proof: (i) This follows since if $v \in V^o(P)$, then both ψ_Λ and ψ_{Λ^T} map v into the leaf-set of H. Therefore, $\psi_{root}(v) \neq \psi_\Lambda(v)$ and $\psi_{root}(v) \neq \psi_{\Lambda^T}(v)$.

(ii) The proof of the stated equalities is in two steps. In the first step, we show that $diam(\mathcal{C}(P, H, \phi), d_{RF}) \leq |V(H)| - 1$ must always hold. In the second step, we use a "proof by contradiction" strategy to show that there cannot exist a non-root vertex v of H such that the set of leaves below the vertex in $H(\psi_\Lambda)$ that v corresponds to equals the set of leaves below the vertex in $H(\psi_{\Lambda^T})$ that v corresponds to. Thus every one of the $|V(H)| - 1$ non-root interior vertices of H must contribute to the Robinson-Foulds distance between ψ_Λ and ψ_{Λ^T}. This will complete the proof since, by symmetry, it implies that any non-trivial cluster in $H(\psi_\Lambda)$ (which must be induced by some non-root vertex in $V(H)$) is not contained in the set of clusters induced by $H(\psi_{\Lambda^T})$ (and vice-versa), and $H(\psi_\Lambda)$ and $H(\psi_{\Lambda^T})$ both induce $|V(H)| - 1$ non-trivial clusters.

To see Step 1, observe that since for any $\psi, \psi' \in \mathcal{C}(P, H, \phi)$ the set of trivial clusters in $H(\phi)$ and $H(\phi')$ coincide and $|V^o(H(\psi))| = |V(H)| = |V^o(H(\psi'))|$, it follows that $diam(\mathcal{C}(P, H, \phi), d_{RF}) \leq |V(H)| - 1$.

To see Step 2, suppose $v \in V(H) - \{\rho_H\}$. Let $Y = L(v)$ denote the set of leaves in H that are below v in H. Note that v gives rise to some interior vertex u (respectively w) in $H(\psi_\Lambda)$ (respectively $H(\psi_{\Lambda^T})$).

Let $L(u)$ (respectively $L^T(w)$) denote the (necessarily non-empty) subset of $L(H(\psi_\Lambda))$ below u (respectively $L(H(\psi_{\Lambda^T}))$ below w). Note that $Y = L(H) \cap L(u) = L(H) \cap L^T(w)$ i.e., the set of leaves of H below v coincides with the set leaves below the vertex in $H(\psi_\Lambda)$ corresponding to v and also with the set of leaves below

the vertex in $H(\psi_\Lambda^T)$ corresponding to v. It follows that if $L(u)$ is equal to $L^T(w)$ then u and w must both correspond to v.

We now claim that this cannot be the case, i.e., if u and w are the vertices in $H(\psi_\Lambda)$ and $H(\psi_{\Lambda^T})$ respectively that both correspond to v, then $L(u) \neq L^T(w)$. To establish this claim, let Z be the subset of $L(P)$ consisting of those leaves x with $\phi(x) \in Y$ i.e., Z is the set of leaves of P that are mapped to a vertex below v under the leaf-map. Note that since $Y \neq L(H)$ and ϕ is surjective, $Z \neq L(P)$. Put differently, since there exists a leaf of H that is not below v, there must exist a leaf of P that is not mapped to a vertex below v under the leaf-map.

Let A (respectively B) be the subset of paths in Λ (respectively Λ^T), whose set of leaves is equal to Z. We will show that the set of vertices crossed by a path in A cannot coincide with the set of vertices crossed by a path in B, i.e., that

$$\bigcup_{\gamma \in A} V(\gamma) \neq \bigcup_{\gamma' \in B} V(\gamma') \tag{1}$$

holds. This will complete the proof of the claim since $L(u) = \{\psi_\Lambda(s) : s \in \bigcup_{\gamma \in A} V(\gamma)\} \cup Y$ and $L^T(w) = \{\psi_{\Lambda^T}(s) : s \in \bigcup_{\gamma' \in B} V(\gamma')\} \cup Y$ as it means that a vertex in Y or on a path in A is mapped under ψ_Λ to a vertex below the vertex in $H(\psi_\Lambda)$ corresponding to v. Similarly, a vertex in Y or on a path in B is mapped under ψ_{Λ^T} to a vertex below the vertex in $H(\psi_{\Lambda^T})$ corresponding to v.

For the purpose of contradiction, suppose equality holds in (1). Let t be contained in $\bigcup_{\gamma \in A} V(\gamma)$ which is maximal with respect to $\preceq_P$.

If $t \neq \rho_P$, then $par(t)$ must be contained in a path in A or a path in B. Without loss of generality, $par(t) \in \bigcup_{\gamma \in A} V(\gamma)$. Hence, $par(t) \in \bigcup_{\gamma \in A} V(\gamma) = \bigcup_{\gamma' \in B} V(\gamma')$, which contradicts the choice of t.

So, suppose $t = \rho_P$. Then since every vertex in P has outdegree two and, by

construction, one of its outgoing edges must be contained in a path in A and the other in a path in B, it follows that $V(P) = \bigcup_{\gamma \in A} V(\gamma)$ as $\bigcup_{\gamma \in A} V(\gamma) = \bigcup_{\gamma' \in B} V(\gamma')$. But then $L(P) = Z$, which is impossible. ∎

Informally speaking Theorem 2 says that the discrete distance between any two reconciliations in $\mathcal{C}(P, H, \phi)$ can be no more than the number of vertices in P, and this distance is attained for the reconciliations ψ_{root} and ψ_Λ. Furthermore, the Robinson-Foulds distance between two reconciliations in $\mathcal{C}(P, H, \phi)$ can be no more than the number of vertices in H minus 1, and this distance is attained for ψ_Λ and ψ_{Λ^T}.

Determining the diameters of the triplet and path distances on $\mathcal{C}(P, H, \phi)$ appears to be a difficult problem (which is also the case for the corresponding metrics on phylogenetic trees – cf. (Steel and Penny 1993)). In practice, we normalise the values of these distances. In the case of the triplet distance, we use for $p = |V(P)|$ and $l = |L(H)|$ the quantity $\binom{p+l}{3} - \binom{l}{3}$ which counts for a reconciliation $\psi \in \mathcal{C}(P, H, \phi)$ the number of triplets displayed by $H(\psi)$. In the case of the path distance, we use the quantity $|V^o(P)| \cdot \max_{l,l' \in L(H)} D_H(l, l')$. Note that both quantities can be shown to be upper bounds for the respective diameters. It would also be of interest to determine the diameter for the edit distance on $\mathcal{C}(P, H, \phi)$ but again this appears to be a challenging problem.

Interestingly, although this is not needed in our computational results, we can also determine diameters for our metrics on the set $\mathcal{R}(P, H, \phi)$ of reconciliations without host switches (Theorem 3). To do this, we first define the reconciliation ψ_{lca} in $\mathcal{R}(P, H, \phi)$, given by setting $\psi_{lca}(v) = m(v)$ for every interior vertex v of P (the so-called lca reconciliation). We also let $\min(P, H, \phi)$ denote the set of non-root interior vertices u of H such that $u = m(v)$ for some vertex v of P and there is no interior vertex v' of P with $m(v')$ strictly below u in H.

Informally speaking, Theorem 3 says that the diameters of the discrete/path/Robinson-Foulds distances on $\mathcal{R}(P, H, \phi)$ are all given by taking the distance

between ψ_{root} and ψ_{lca}.

Theorem 3. *Given (P, H, ϕ) we have*

(i)

$$\begin{aligned} diam(\mathcal{R}(P,H,\phi), d_{disc}) &= d_{disc}(\psi_{root}, \psi_{lca}) \\ &= |\{v \in V^o(P) \,:\, m(v) \neq \rho_H\}|. \end{aligned}$$

(ii)

$$\begin{aligned} diam(\mathcal{R}(P,H,\phi), d_{path}) &= diam(\mathcal{R}(P,H,\phi), d_{ed}) = d_{path}(\psi_{root}, \psi_{lca}) \\ &= \sum_{v \in V^o(P)} D_H(\rho_H, m(v)). \end{aligned}$$

(iii)

$$\begin{aligned} diam(\mathcal{R}(P,H,\phi), d_{RF}) &= d_{RF}(\psi_{root}, \psi_{lca}) \\ &= |V(H)| - 1 - \sum_{u \in \min(P,H,\phi)} (|V(H_u)| - 1). \end{aligned}$$

Proof: Statements (i) and (ii) follow immediately from the following observation. Suppose $\psi, \psi' \in \mathcal{R}(P, H, \phi)$. Then by Lemma 1, for all $v \in V(P)$, the vertices $\psi(v), \psi'(v)$ are both contained in the path $A(v)$ in H, which has endpoints ρ_H and $m(v)$. Hence the maximum possible distance between $\psi(v)$ and $\psi'(v)$ in H (relative to the path distance D_H in H) is the length of the path $A(v)$, and this is achieved in case $\{\psi(v), \psi'(v)\} = \{\rho_H, m(v)\}$. But this is the case for all $v \in V(P)$ when $\psi = \psi_{root}$ and $\psi' = \psi_{lca}$.

(iii) First note that if $\psi \in \mathcal{R}(P, H, \phi)$, then $|Cl(H(\psi))| = |V(H)| + |L(H)| + |V(P)|$,

where $Cl(H(\psi))$ denotes the set of clusters in $H(\psi)$. Hence, for $\psi, \psi' \in \mathcal{R}(P, H, \phi)$,

$$\begin{aligned} 2d_{RF}(\psi, \psi') &= |Cl(H(\psi))| + |Cl(H(\psi'))| - 2|Cl(H(\psi)) \cap Cl(H(\psi'))| \\ &\leq 2(|V(H)| + |L(H)| + |V(P)|) - \\ &\quad 2[|L(H)| + |V(P)| + 1 + \sum_{u \in \min(P,H,\phi)} |V(H_u) - \{\rho_{H_u}\}|], \end{aligned}$$

since every trivial cluster and every cluster induced by a vertex w in $V(H_u) - \{\rho_{H_u}\}$ considered as a vertex in $H(\psi)$ and also in $H(\psi')$ for $u \in \min(P, H, \phi)$, must be contained in both $H(\psi)$ and $H(\psi')$. Moreover, it is straightforward to check that equality holds in case $\psi = \psi_{root}$ and $\psi' = \psi_{lca}$. Statement (iii) now follows immediately. ∎

It is interesting to note that we can determine the diameter of the path distance on $\mathcal{R}(P, H, \phi)$ but not on $\mathcal{C}(P, H, \phi)$. This is because by Lemma 1, we have more control on the image of a map in $\mathcal{R}(P, H, \phi)$ as compared to one that lies in $\mathcal{C}(P, H, \phi)$. As with $\mathcal{C}(P, H, \phi)$ it also appears to be difficult to determine the diameter of the triplet distance on $\mathcal{R}(P, H, \phi)$.

Computational results

To evaluate the performance of our metrics on biological datasets, we obtained 8 pairs of host and parasite trees from the literature. The choice of datasets was driven by the availability of the data in public databases, and the desire to test the metrics on a wide variety of datasets both in terms of size and the topology of the trees. We therefore chose data where (a) both host and parasite tree have relatively small size, (b) the host tree is much smaller than the parasite tree, (c) the host and parasite tree both have medium size, and (d) the host and parasite trees are both large. We present a summary of the datasets that we used in Table 1.

Dataset name	Organisms involved	Reference	Type of dataset	Number of hosts	Number of parasites
FE	*Formicidae & Eucharitidae*	Murray *et al.* (2013)	(a)	4	5
CP	*Cichlidae & Platyhelminthes*	Mendlová *et al.* (2012)	(b)	6	29
PMP	*Pelican & Lice*	Hughes *et al.* (2007)	(c)	18	18
RH	*Rodents & Hantaviruses*	Ramsden *et al.* (2009)	(d)	34	42
EC	*Encyrtidae & Coccidae*	Deng *et al.* (2013)	(a)	7	10
GL	*Gopher & Lice*	Hafner and Nadler (1988)	(a)	8	10
SC	*Seabirds & Chewing Lice*	Paterson *et al.* (2003)	(c)	11	14
SFC	*Smut Fungi & Caryophillaceus plants*	Refregier *et al.* (2008)	(c)	15	16

Table 1: For each dataset used we present its name and reference, its type, number of leaves in the host tree, and number of leaves in the parasite tree.

Full results and details for the remaining datasets that we considered are presented in Supplementary Material II available from Dryad at https://doi.org/10.5061/dryad.g38g40b. Note that the case where the host tree is much larger than the parasite tree is not interesting for our study since we assume that every host has a parasite.

The phylogenetic trees for these datasets can be found in Supplementary Material II. We computed reconciliations between these trees using the EUCALYPT software (Donati *et al.* 2015). The dataset FE is relatively small and there are only 215 possible reconciliations, hence for this dataset, our analysis was performed on the whole space of reconciliations. The space of possible reconciliations for the other trees grows extremely quickly; for example for the dataset EC in which the host and parasite tree have 7 and 10 leaves, respectively, the total number of reconciliations equals 34,359. Hence, for each of the remaining datasets, 200 reconciliations were uniformly sampled from the whole space of possible reconciliations using the method implemented in (Donati *et al.* 2015) with the -random parameter and the cost of each one of the events set to 0. We implemented the d_{RF}, d_{disc}, d_{path} and d_{tr} distances in python; the code is available from the website mentioned in the abstract. To manipulate phylogenetic trees we used the ETE toolkit for python (Huerta-Cepas *et al.* 2016).

Distributions and Correlations

For every pair of reconciliations in each dataset we calculated the metrics $d_{disc}, d_{path}, d_{tr}, d_{RF}$ in order to produce a distribution of pairwise values. We did not consider the edit distance d_{ed} as we do not know of an efficient algorithm for its computation. All metrics were normalised by the diameters given above. The distribution of values for the pairwise distances for the elements in each of the four datasets are depicted in Fig. 3.

In general, the results indicate that d_{disc} and d_{RF} have similar distributions, which are somewhat different from the d_{tr} and d_{path} distributions, which are in turn quite similar. Indeed, when the d_{RF} values are high, then so are the d_{disc} values, and the same holds for the d_{path} and d_{tr} metrics. However, note that in the CP dataset the values assumed by the discrete distance are smaller that the ones for the Robinson-Foulds distance. This is because the parasite tree is much larger than the host tree and this influences the normalisation that we use for these distances. It is also worth noting that the distribution of the d_{RF} metric appears shifted to the right, implying that most of the pairs of reconciliations are far apart in the Robinson-Foulds distance. This is quite similar to the behaviour of the Robinson-Foulds distance for phylogenetic trees (Steel and Penny 1993).

The distributions suggest that the d_{path} and d_{tr} metrics are better at discriminating between reconciliations than the d_{RF} and d_{disc} metrics since they attain a larger spread of values. Hence, it might be better to use the d_{path} metric especially as this metric can be computed quite efficiently. Finally, note that the triplet distance attains very small values, but this is because we are normalising using a value that is probably much larger than the diameter of d_{tr}. Thus, finding a better upper bound for the diameter of d_{tr} would be useful.

To get a better understanding on how the metrics compare with one another on real data, we also produced pairwise scatter plots for each of the datasets. Since the plots have many overlapping points which can obscure the density of the data in a particular region,

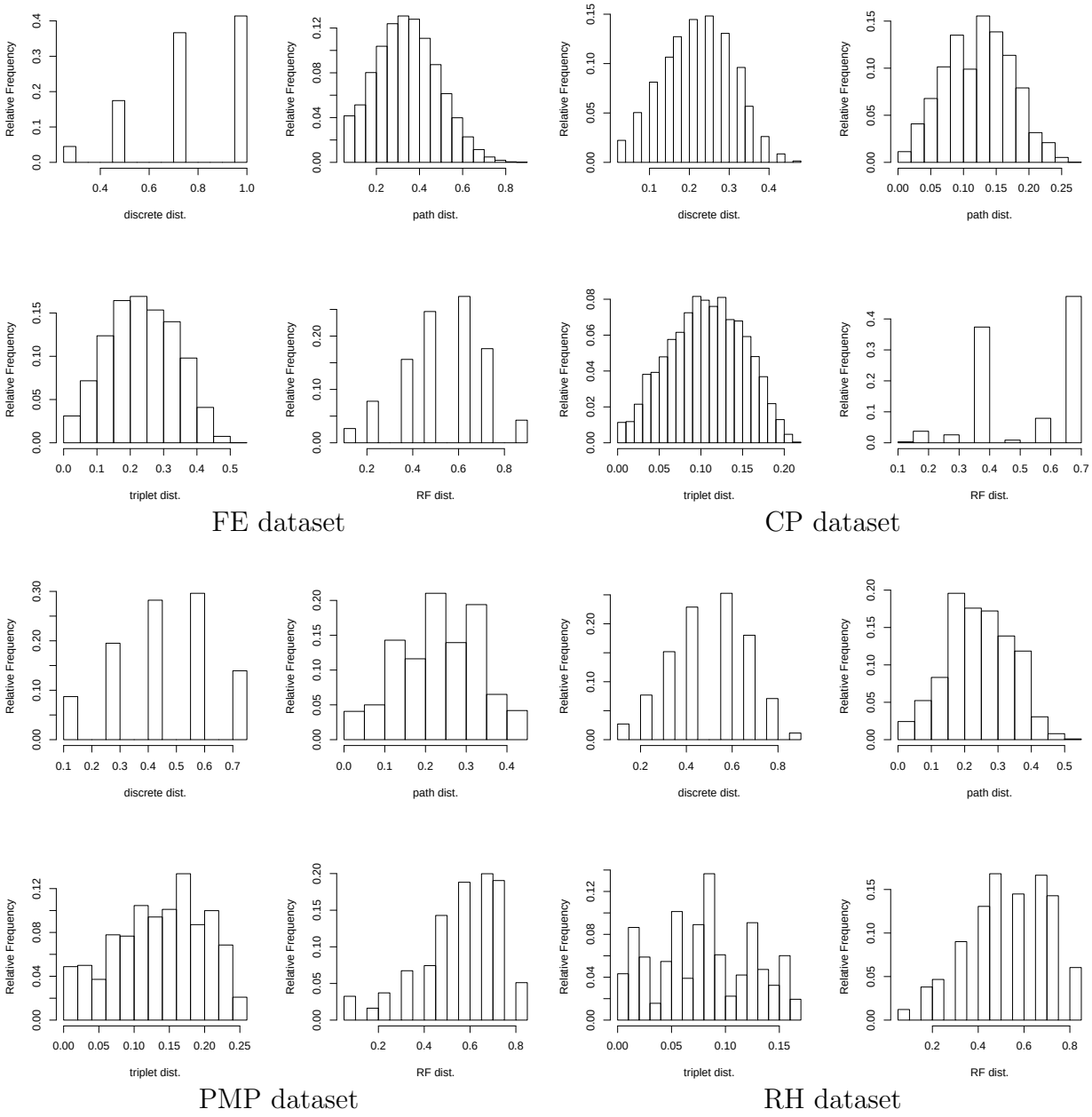
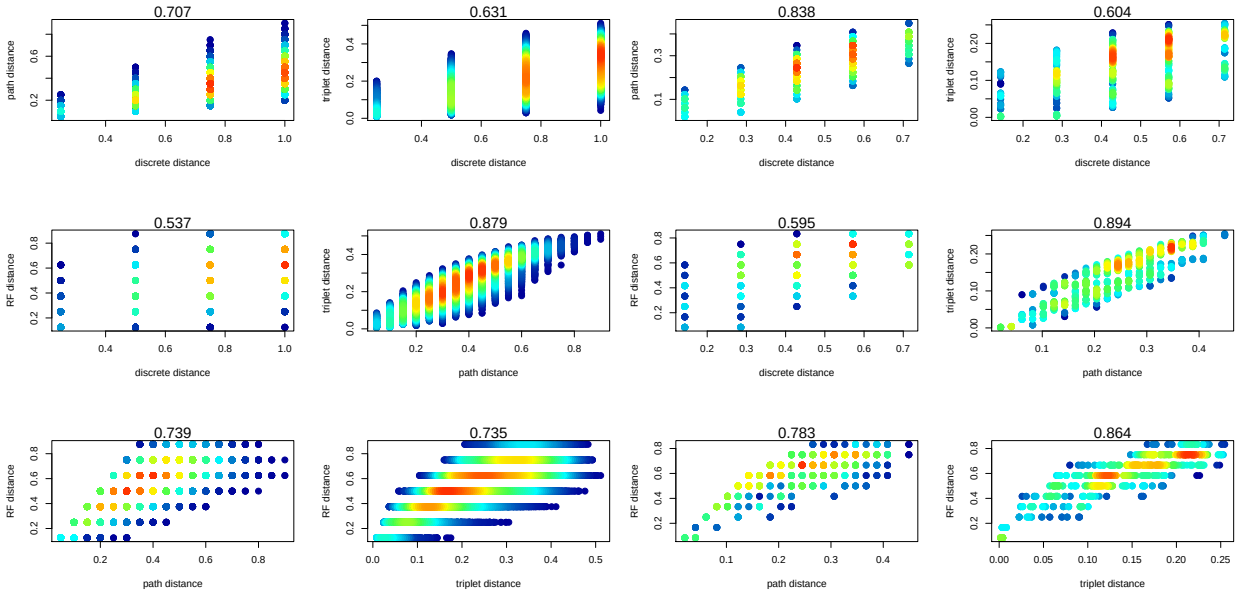


Figure 3: The relative frequency of the values of the normalised pairwise distances between reconciliations, for the FE, CP, PMP and RH datasets (in top left, top right, bottom left and bottom right). For the dataset FE the whole space of 215 reconciliations is considered, while for the other datasets we present a sample of 200.

we used a heat-map colouration in our plots to indicate the number of points that are overlapping (in the online version, a clear black means low density and red means high

density). In Fig. 4 we present plots for the FE and PMP datasets; the other results are similar and may be found in the Supplementary Material II.

In the plots, we observe that there is generally a good correlation between each of the metrics, except for d_{RF} which does not seem to correlate well with any other metric. Perhaps not surprisingly the highest correlation is observed between d_{path} and d_{disc}. We also observe that d_{path} correlates strongly with d_{tr}. This again suggests that the path distance could be a good alternative to the triplet distance.



FE dataset PMP dataset

Figure 4: Scatter plot for the normalised distances on the FE (left) and PMP (right) datasets. The value of the Pearson correlation coefficient is given at the top of each panel.

Multidimensional scaling

Multidimensional scaling (MDS) is a useful technique for visualising a distance matrix (Cox and Cox 2000). It essentially works by trying to represent the matrix by a collection of points in 2- or 3-dimensional Euclidean space so that the Euclidean distances

between the points match the original distances as closely as possible. This technique has been used to visualise phylogenetic tree space in, for example, Hillis *et al.* (2005); Jombart *et al.* (2017).

We investigated MDS as a possible way to visualise reconciliation spaces. We calculated the distances between all pairs of reconciliations and obtained 2-dimensional MDS representation using the R package (Team 2013). In the online version, we also assigned a colour to each point corresponding to the cost of the reconciliation represented by the point (see the legend in the top right of each graphic for the correspondence between colours and cost values), to obtain some idea of how optimal/suboptimal reconciliations are distributed in reconciliation space. The cost is computed relative to a model in which events are assigned a certain cost, and is represented by a cost vector. We used the two cost vectors that are most commonly used in the literature: $c_c = 0$; $c_d = c_s = c_l = 1$ and $c_c = 0$; $c_d = c_s = 1$; $c_l = 2$, where c_c, c_d, c_s and c_l are the costs for cospeciation, duplication, speciation and loss, respectively (see (Donati *et al.* 2015) for more details concerning these costs and the underlying model). Then for each of the reconciliations that were uniformly sampled from the whole space, the cost is calculated under the two cost vectors chosen. In this way, under a given cost vector, some of these reconciliations may have an optimal cost or a slightly higher cost compared to the optimal value.

As an illustration of the plots that we obtained, in Fig. 5 we present the MDS plots for the PMP dataset for the two costs that we considered; the results for the other datasets may be found in Supplementary Material II. Note that the x- and y-axis are the coordinates of the 2-dimensional Euclidean space which is being used to represent the distances between reconciliations. As might be expected, since we picked a random sample of reconciliations for all but the FE dataset, in general we found that the reconciliations were quite spread out in most of the plots. The exception to this were the plots involving the d_{RF} metric, where except for the FE and GL datasets, we obtained a small number of

clusters. This is probably related to the fact that the d_{RF} distributions took on much fewer values.

Interestingly, we found that the optimal or close to optimal reconciliations were quite spread out throughout the MDS plots in general. This tends to suggest that the landscape of the reconciliation spaces that we consider are somewhat rugged, with clusters of optimal and close to optimal reconciliations being rather spread out instead of forming distinct peaks. We should mention however, that for the d_{path} metric in the MDS plot for the RH dataset with cost $c_c = 0$, $c_d = 1$, $c_s = 2$, and $c_c = 1$, we observed a rather tight clustering of close to optimal cost reconciliations, indicating that there may be a rather sharp peak surrounding the optimal solution. A similar peak cluster can be observed in the MDS plot for the SC dataset with the same cost vector, although in this example, there is no corresponding peak in the MDS plot for the triplet distance. This indicates that the choice of distance could be key when exploring spaces of reconciliations.

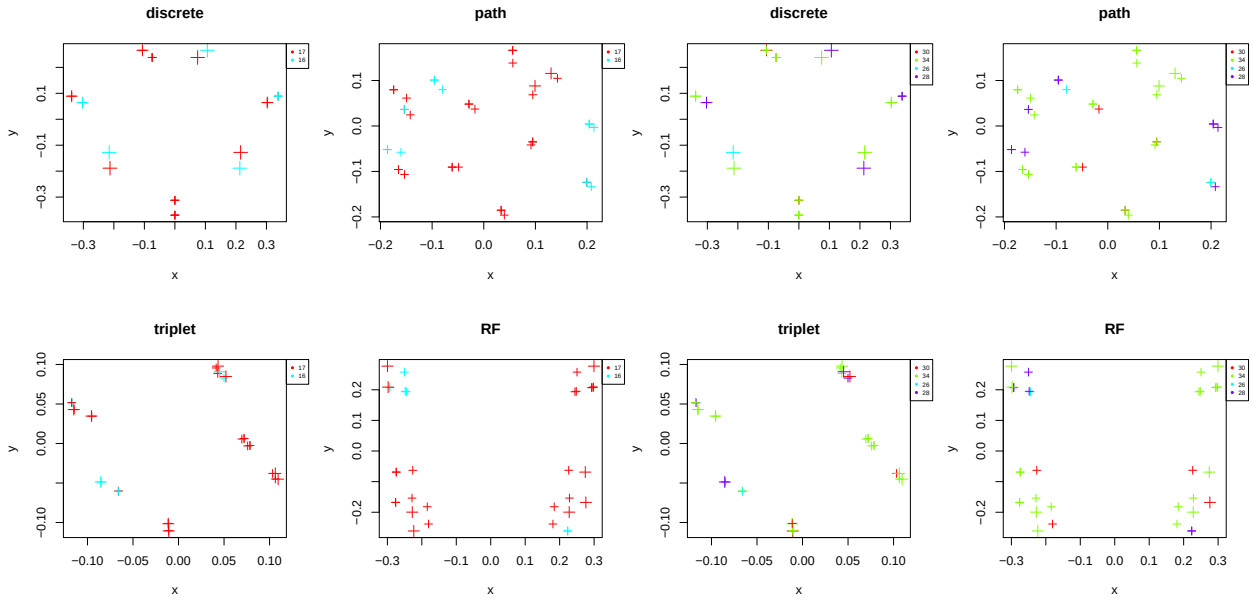


Figure 5: MDS for the PMP dataset, using a sample of size 200. The cost of a reconciliation is represented in the online version by the colour of the corresponding point according to the legend in the top right. Left panel uses cost vector $c_c = 0$, $c_d = 1$, $c_s = 1$, $c_l = 1$ (optimal cost is 14) and the right panel uses the cost vector $c_c = 0$, $c_d = 1$, $c_s = 2$, and $c_l = 1$ (optimal cost is 26).

Concluding comments

In this paper, we have presented various metrics on reconciliation spaces, and also given formulae for some of their diameters. We have also investigated their properties in practise by computing their distributions and comparing them for a variety of real biological datasets. Based on our analysis, we would recommend the use of the path distance amongst those metrics that we have considered for most applications. Our computational results also indicate that our distances used in combination with multidimensional scaling (MDS) could be a useful for tool for exploring and visualising reconciliation spaces.

We should note that in practice, we do not advocate the exclusive use of the metrics that we have considered in this paper. Indeed, just as with tree-spaces, there may be other metrics which could be useful for understanding collections of reconciliations based on different biologically motivated questions. Even so, the framework that we have presented for exploring reconciliation spaces using MDS is potentially useful for any choice of metric, and it could be interesting to explore further whether other metrics might be useful in this context.

Note that the reconciliation model that we considered does not allow for widespread parasites, that is, for parasites associated with more than one host, and it would be interesting to see if our results could be extended to accommodate this possibility. Some models of reconciliations already permit widespread parasitism, such as those in Merkle *et al.* (2010) and Jane 4 (Conow *et al.* 2010) which allow a restricted version of widespread parasitism. Even so, there is no commonly accepted way to deal with widespread parasitism, and therefore extending our results to include this possibility would at least require developing ways to generalise the metrics that we have considered.

From a theoretical standpoint, we have some open questions. First, it would be of

interest to obtain better bounds for the triplet distance on $\mathcal{C}(P, H, \phi)$ and $\mathcal{R}(P, H, \phi)$ and also the diameter for the path distance on $\mathcal{C}(P, H, \phi)$ for a host-parasite triple (P, H, ϕ). This might also involve finding good lower bounds for these metrics, which could be potentially derived by a more careful consideration of the arguments presented in Theorem 3. Second, although it follows from our results that we can efficiently compute the edit distance on $\mathcal{R}(P, H, \phi)$ (since on this set the edit distance and the path distance coincide), it would be interesting to know the complexity of computing the edit distance on $\mathcal{C}(P, H, \phi)$. We suspect that in general it will be NP-hard to compute. Third, as it can be very expensive to compute all pairwise distances for large sets of reconciliations, it could be useful to explore and implement more efficient algorithms for their computation than the ones we have made available (for example, our implementation of the triplet distance might be sped up using techniques given in Brodal *et al.* (2013)). Fourth, in general we have not specifically considered spaces of time-feasible reconciliations (ones in which no contradictory temporal scenarios arise – see for example (Donati *et al.* 2015)); it would be interesting to understand how our analyses might apply to these spaces.

In the MDS experiments, we generally found that optimal or nearly optimal reconciliations were quite spread out across the space of reconciliations. This is quite different from, for example, spaces of phylogenetic trees which have been found to have well-defined peaks of suboptimal trees forming around an optimal tree (Hillis *et al.* 2005). This seems to indicate that the spaces of reconciliations form a rugged landscape; it could be of interest to investigate this further using techniques from combinatorial landscape analysis (see for example (Charleston 1995) and (Reidys and Stadler 2002)). From a practical point of view, this means that some care should be taken in considering only a single optimal solution when trying to compute reconciliations. To explore this further, it could be of interest to perform simulations (for example using the ALF tool (Dalquen *et al.* 2012)) to study the effect that tree topologies and costings might have on metric

distributions and MDS plots.

As we have seen in our results, there may be several optimal solutions in general (or at least many close to optimal solutions), and these could be quite different from one another. Thus there is a need for new approaches to either reduce the number of optimal solutions (for example by refining the models) or to maybe sample and summarise properties of collections of optimal solutions (for example by developing consensus approaches as in Haack *et al.* (2018); Huber *et al.* (2018); Nguyen *et al.* (2013)). We envisage that the techniques that we developed in this paper will provide a useful addition to such investigations.

Supplementary Material

Supplementary Material I and II are available from the Dryad Digital Repository at **https://doi.org/10.5061/dryad.g38g40b**

Acknowledgements

All authors thank the Royal Society for its support through their International Exchanges Scheme. Also they thank the anonymous reviewers and the Associate Editor M. Charleston for their helpful comments and suggestions.

*

References

Bansal, M. and Alm, E. Kellis, M. 2012. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, 28(12): i283–i291.

Bansal, M., Alm, E., and Kellis, M. 2012. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, 28(12): i2839.

Bansal, M., Mukul, S., Alm, E., and Kellis, M. 2013. Reconciliation revisited: Handling multiple optima when reconciling with duplication, transfer, and loss. *J. Comp. Biol.*, 20(10): 738–754.

Billera, L., Holmes, S., and Vogtmann, K. 2001. Geometry of the space of phylogenetic trees. *Adv. App. Math*, 27: 733–767.

Brodal, G., Fagerberg, R., Mailund, T., S., P. C., and Sand, A. 2013. Efficient algorithms for computing the triplet and quartet distance between trees of arbitrary degree. *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1814–1832.

Chan, Y., Ranwez, V., and Scornavacca, C. 2015. Exploring the space of gene/species reconciliations with transfers. *J. Math. Biol.*, 71: 1179–1209.

Charleston, M. 1995. Toward a characterization of landscapes of combinatorial optimization problems, with special attention to the phylogeny problem. *J. Comp. Biol.*, 2(3): 439–50.

Charleston, M. A. 1998. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math. Biosci.*, 149(2): 191–223.

Charleston, M. A. 2003. Recent results in cophylogeny mapping. *Adv. Parasit.*, 54: 303–30.

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. 2010. Jane: a new tool for the cophylogeny reconstruction problem. *Alg. Mol. Biol.*, 5(16): 10 pages.

Cox, T. and Cox, M. 2000. *Multidimensional Scaling (2nd Edition)*. Chapman and Hall/CRC.

Dalquen, D. A., Anisimova, M., Gonnet, G. H., and Dessimoz, C. 2012. Simulation framework for genome evolution. *Mol. Biol. Evol.*, 29(4): 1115–1123.

Deng, J., Yu, F., Li, H.-B., Gebiola, M., Desdevises, Y., Wu, S.-A., and Zhang, Y.-Z. 2013. Cophylogenetic relationships between Anicetus parasitoids (Hymenoptera: Encyrtidae) and their scale insect hosts (Hemiptera: Coccidae). *BMC Evol. Biol.*, 13(1): 1–11.

Donati, B., Baudet, C., Sinaimeri, B., Crescenzi, P., and Sagot, M.-F. 2015. Eucalypt: efficient tree reconciliation enumerator. *Alg. Mol. Biol.*, 10(1): 3.

Doyon, J., Chauve, C., and Hamel, S. 2009. Space of gene/species tree reconciliations and parsimonious models. *J. Comp. Biol.*, 16(10): 1399–1418.

Doyon, J., Ranwez, V., Daubin, V., and Berry, V. 2011. Models, algorithms and programs for phylogeny reconciliation. *Brief. Bioinform.*, 12(5): 392–400.

Doyon, J., Hamel, S., and Chauve, C. 2012. An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *IEEE/ACM Trans. Comp. Bio. Bioinf.*, 9(1): 26–39.

Eulenstein, O., Huzurbazar, S., and Liberles, D. 2010. Reconciling phylogenetic trees. In K. Dittmar and D. Liberles, editors, *Evolution after gene duplication*, pages 185–206. Wiley-Blackwell.

Haack, J., Zupke, E., Ramirez, A., Wu, Y.-C., and Libeskind-Hadas 2018. Computing the diameter of the space of maximum parsimony reconciliations in the Duplication-Transfer-Loss model. *IEEE/ACM Trans. Comput. Bio. Bioinf*.

Hafner, M. and Nadler, S. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, 332: 258–259.

Hillis, D., Heath, T., and St John, K. 2005. Analysis and visualization of tree space. *Syst. Biol.*, 54(3): 471–482.

Huber, K. T., Spillner, A., and Moulton, V. 2011. Metrics on multilabeled trees: interrelationships and diameter bounds. *IEEE/ACM Trans. Comp. Bio. Bioinf.*, 8(4): 1029–1040.

Huber, K. T., Moulton, V., Sagot, M., and Sinaimeri, B. 2018. Geometric medians in reconciliation spaces. *Information Processing Letters*, 136: 96–101.

Huerta-Cepas, J., Serra, F., and Bork, P. 2016. Ete 3: Reconstruction, analysis and visualization of phylogenomic data. *Mol. Biol. Evol.*, 33(6): 16351638.

Hughes, J., Kennedy, M., Johnson, K. P., Palma, R. L., and Page, R. D. 2007. Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and Pectinopygus lice. *Syst. Biol.*, 56(2): 232–251.

Jombart, T., Kendall, M., AlmagroGarcia, J., and Colijn, C. 2017. Treespace: statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.*, 17(6): 1385–1392.

Kendall, M. and Colijn, C. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol. Biol. Evol.*, 33(10): 2735–2743.

Libeskind-Hadas, R., Wu, Y., Bansal, M., and Kellis, M. 2014. Pareto-optimal phylogenetic tree reconciliation. *Bioinformatics*, 30(12): i87–95.

Mendlová, M., Desdevises, Y., Civáňová, K., Pariselle, A., and Šimková, A. 2012. Monogeneans of West African cichlid fish: evolution and cophylogenetic interactions. *PLoS One*, 7(5): e37268.

Merkle, D. and Middendorf, M. 2005. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences*, 123: 277–299.

Merkle, D., Middendorf, M., and Wieske, N. 2010. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC bioinformatics*, 11(1): 560.

Moulton, V., Zuker, M., Steel, M., Pointon, R., and Penny, D. 2000. Metrics on RNA secondary structures. *J. Comp. Biol.*, 7: 277–292.

Murray, E. A., Carmichael, A. E., and Heraty, J. M. 2013. Ancient host shifts followed by host conservatism in a group of ant parasitoids. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1759): 20130495.

Nakhleh, L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution*, 28(12): 719–28.

Nguyen, T., Ranwez, V., Berry, V., and Scornavacca, C. 2013. Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods. *PloS one*, 8(10): e73667.

Page, R. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, 43(1): 58–77.

Page, R. D. and Charleston, M. A. 1998. Trees within trees: phylogeny and historical associations. *Trends in Ecology & Evolution*, 13(9): 356–9.

Paterson, A. M., Ricardo, L. P., and Gray, R. D. 2003. Drowning on arrival, missing the boat, and x-events: how likely are sorting events? In R. Page, editor, *Tangled trees: Phylogeny, cospeciation, and coevolution*, pages 287–309. University Chicago Press.

Ramsden, C., Holmes, E. C., and Charleston, M. A. 2009. Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Mol. Biol. Evol.*, 26(1): 143–153.

Refregier, G., Le Gac, M., Jabbour, F., Widmer, A., Shykoff, J., Yockteng, R., Hood, M., and Giraud, T. 2008. Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. *BMC Evol. Biol.*, 8(1): 100.

Reidys, C. and Stadler, P. 2002. Combinatorial landscapes. *SIAM REVIEW, Society for Industrial and Applied Mathematics*, 44(1): 3–54.

Ronquist, F. 1995. Reconstructing the history of host-parasite associations using generalised parsimony. *Cladistics*, 11(1): 73–89.

Ronquist, F. 2002. Parsimony analysis of coevolving species associations. In P. R.D.M., editor, *Tangled trees: Phylogeny, cospeciation and coevolution*, pages 22–64. University of Chicago Press.

Scornavacca, C., Paprotny, W., and Berry, V.and Ranwez, V. 2013. Representing a set of reconciliations in a compact way. *J. of Bioinf. Comp. Biol.*, 11(02): 1250025.

Steel, M. and Penny, D. 1993. Distributions of tree comparison metrics–some new results. *Syst. Biol.*, 42(2): 126–141.

Stolzer, M., Lai, H., Xu, M. amd Sathaye, D., Vernot, B., and Durand, D. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18): i409–15.

Team, R. C. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tofigh, A., Hallett, M., and Lagergren, J. 2011. Simultaneous identification of duplications and lateral gene transfer. *IEEE/ACM Trans. Comp. Bio. Bioinf.*, 8(2): 517–535.

Wu, T. and Zhang, L. 2012. Structural properties of the reconciliation space and their applications in enumerating nearly-optimal reconciliations between a gene tree and a species tree. *BMC Bioinformatics*, 12((Suppl 9)): S7.